BIOMETRICS 57, 779–786
September 2001

A Bivariate Cure-Mixture Approach for Modeling Familial Association in Diseases

Nilanjan Chatterjee

Statistical Research and Applications Section, Biostatistics Branch,
Division of Cancer Epidemiology and Genetics, National Cancer Institute,
6120 Executive Boulevard, EPS 8038, Rockville, Maryland 20852, U.S.A.
email: chattern@mail.nih.gov

and

Joanna Shih

Biometric Research Branch, National Cancer Institute, Rockville, Maryland, U.S.A.

SUMMARY. For modeling correlation in familial diseases with variable ages at onset, we propose a bivariate model that incorporates two types of pairwise association, one between the lifetime risk or the overall susceptibility of two individuals and one between the ages at onset between two susceptible individuals. For estimation, we consider a two-stage estimation procedure similar to that of Shih (1998, *Biometrics* **54,** 1115–1128). We evaluate the properties of the estimators through simulations and compare the performance with that from a bivariate survival model that allows correlation between ages at onset only. We apply the methodology to breast cancer using the kinship data from the Washington Ashkenazi Study. We also discuss potential applications of the proposed method in the area of cure modeling.

KEY WORDS: Copula models; Correlated failure times; Cure models; Mixture models; Semiparametric models.

1. Introduction

Consider two types of expression for a disease, the incidence and the age at onset of disease for the diseased individuals. Risk models for overall susceptibility (lifetime risk) that consider only the first expression by treating the disease as a binary trait of being affected or unaffected can produce misleading results because, for individuals without the disease, due to incomplete follow-up, it is often not known whether they will eventually develop the disease. On the other hand, models for survival analysis typically assume that everybody has the same susceptibility to the disease and will eventually develop it if followed up sufficiently long. These models may not properly describe the disease risk since risk factors such as genetic predispositions can conceivably make one group of individuals more susceptible to the disease than other groups. In models where both types of expressions are considered, the effect of a covariate or a risk factor can act on either the overall susceptibility or the age at onset or both. In segregation analysis, e.g., where one tests for a major gene effect using familial disease data, Elston and Yelverton (1975) proposed models where the effect of a gene can be manifested either in the susceptibility or the age at onset of the disease. Both types of models are commonly used in practice (cf., Claus, Risch, and Thompson, 1991), and it is debatable which model will be more appropriate in a particular situation. Although it has often been assumed in many studies of breast cancer that the effect of the gene is manifested as an earlier age at onset, Ciske et al. (1996) found that, for some families, the effect of a putative susceptibility gene can be better modeled by representing its risk in terms of overall susceptibility to the disease rather than a shift in the age-at-onset distribution.

In this article, we consider the problem of modeling familial correlation of diseases. Standard modeling approaches for correlated binary data, such as GEE (Zhao and Prentice, 1990; Liang, Zeger, and Qaqish, 1992), may not be able to account for the age-at-onset information and the possible censoring of the controls in a satisfactory way. Both of these problems can be accounted for by using models for bivariate or multivariate survival data, such as copula or frailty models. These models, however, specify the correlation among individuals only in terms of their ages at onset of the disease. Given that familial correlation of a disease is typically caused by aggregation of various genetic or/and environmental risk factors of the disease and the effect of such factors can be either in the susceptibility or the age at onset, it is plausible that familial correlation, or at least part of it, will be better expressed in terms of the overall susceptibility. To account for such possibilities, in this article, we propose a model for bivariate data that specifies the correlation between overall susceptibility of two individuals and also the correlation between the ages at onset of two susceptible individuals. Random effect models and marginal models are two common approaches for analyzing correlated data. We consider the marginal approach, i.e., the marginal distributions are not affected by the dependency structure. Similar to the spirit of Shih (1998), a two-stage estimation approach allowing for possible censoring is considered for estimation of the parameters of the model.

An application of this research is considered using the real data example from the Washington Ashkenazi Study (WAS) (Struewing et al., 1997). In this study, more than 5000 volunteer Ashkenazi Jews living in the Washington, D.C., area provided blood samples for genotyping of BRCA1/BRCA2 mutations and family history information on breast and some other common cancers, including ages at onset for the first-degree relatives. One primary interest of the study was to estimate the risk of breast cancer among the carriers and noncarriers of the gene mutations. In this article, we will only use the family history data on breast cancer for the first-degree female relatives of the volunteer participants, the interest being estimation of the correlation between susceptibility of the relatives as well as between the ages at onset of the susceptible relatives. Estimating these two types of correlation can be informative in determining whether the influence of familial risk factors for breast cancer is better expressed in terms of overall susceptibility to the disease or by the age at onset of the disease or a combination of both.

The rest of the article is organized as follows. In section 2, we propose the new model and a two-stage estimation approach for the parameters in the proposed model based on a quasi-likelihood of the data. In Section 3, we apply the proposed model to investigating familial association of breast cancer using data from the Washington Ashkenazi Study. In Section 4, we evaluate the performance of the proposed approach using simulation experiments. In Section 5, we discuss other potential applications of our method, particularly in the area of cure modeling. The article is concluded with a discussion on the problem of parameter interpretation when the hypothetical mixture population does not exist.

2. Methods

2.1 *Model*

Use of mixture models has been popular for joint modeling of the overall risk of a disease and the age-at-onset distribution of the diseased individuals (Elston and Yelverton, 1975; Farewell, 1977; Farewell, Math, and Math, 1977; Kuk and Chen, 1992). In what follows, we consider an extension of these mixture models to a bivariate setting. Define an individual to be susceptible if he/she will eventually develop the disease if followed up sufficiently long. For a pair of individuals, $j = 1, 2$, define

$$Y_j = \begin{cases} 1 & \text{if the } j\text{th individual is susceptible} \\ 0 & \text{otherwise} \end{cases}$$

and let T_j^* denote the age at onset for the jth individual when $Y_j = 1$.

With the above notation, we now describe a marginal modeling approach. For $j = 1, 2$, let $\phi_j = \Pr(Y_j = 1)$ and $S_j(t) = \Pr(T_j^* \geq t \mid Y_j = 1)$ describe the marginal distribution of Y_j and the failure time T_j^* for the susceptible individuals, respectively. At this stage of the model, several alternative choices are possible. Depending on specific applications, including the example we consider in Section 3, one may need to assume a common marginal distribution for the two members of the pair, i.e., $\phi_1 = \phi_2$ and $S_1(t) = S_2(t)$ for all t. Also, one may specify parametric forms for $S_1(t)$ and $S_2(t)$ or one can leave them unspecified and treat them nonparametrically.

The next stage of the model involves specifying a dependence structure between the members of the pair. We introduce two types of association, one between the susceptibility to the disease and one between the failure times of two susceptible members. The pairwise odds ratio parameter is a natural choice of the measure of association between Y_1 and Y_2. Let

$$\gamma = \frac{p_{11}p_{00}}{p_{10}p_{01}},$$

where $p_{ij} = \Pr(Y_1 = i, Y_2 = j)$, $i = 0, 1$, $j = 0, 1$. We take a copula modeling approach for specifying the dependency structure between the failure times of two susceptible individuals. Copula models are classes of bivariate survival distributions, specified in terms of the marginal survivor functions and a copula function, a continuous bivariate distribution function on the unit square $[0, 1]^2$ with uniform marginals. Different choices of copula functions impose different association structures between the failure times without changing the marginal distributions.

Let $C_\alpha(u, v)$,$\alpha \in \mathcal{A}$ be a class of distribution functions with uniform margins on $[0, 1]^2$ and assume

$$\Pr(T_1^* \geq t_1, T_2^* \geq t_2 | Y_1 = 1, Y_2 = 1) = C_{\alpha_0}(S_1(t_1), S_2(t_2)) \quad (1)$$

for some $\alpha_0 \in \mathcal{A}$. In our application, we consider three popularly studied models:

(1) Clayton's model (Clayton, 1978),

$$C_\theta(u, v) = \begin{cases} (u^{1-\theta} + v^{1-\theta} - 1)^{1/(1-\theta)}, & \theta > 1, \\ uv, & \theta = 1, \end{cases}$$

(2) Frank's model (Frank, 1979),

$$C_\kappa(u, v) = \begin{cases} \log\left\{1 - \dfrac{(1-\kappa^u)(1-\kappa^v)}{1-\kappa}\right\} / \log\kappa, & 0 < \kappa < 1, \\ uv, & \kappa = 1, \end{cases}$$

(3) Positive stable model (Hougaard, 1986),

$$C_\omega(u, v) = \begin{cases} \exp\left[-\left\{(-\log u)^{1/\omega} + (-\log v)^{1/\omega}\right\}^\omega\right], & 0 < \omega < 1, \\ uv, & \omega = 1. \end{cases}$$

In the above formulae, we have shown the range of the parameter values that corresponds to positive association. In Frank's and Clayton's models, however, one can accommodate negative association, too. Specifically, $\kappa > 1$ and $\theta < 1$ correspond to negative association in Frank's and Clayton's models, respectively.

All the copula parameters, θ, κ, and ω, measure strength of association, but their exact interpretations are different. One way to interpret these parameters is through the cross-ratio function associated with a bivariate survival distribution

$S(t_1, t_2)$ (Clayton, 1978),

$$c(t_1, t_2) = \frac{S(t_1, t_2) D_1 D_2 S(t_1, t_2)}{\{D_1 S(t)\}\{D_2 S(t)\}},$$

where D_j denotes the differential operator $-\partial/\partial t_j$. This function gives a time-dependent measure of association between two failure times and may be interpreted as the ratio of the hazard rate of the conditional distribution of T_1 (T_2) given $T_2 = t_2$ ($T_1 = t_1$) to that of T_1 (T_2) given $T_2 > t_2$ ($T_1 > t_1$) (Oakes, 1989). Oakes showed that, for the general class of Archimedian distributions (Genest and Mckay, 1986), $c(t_1, t_2)$ uniquely characterizes $S(t_1, t_2)$ and depends on (t_1, t_2) only through $v = S(t_1, t_2)$, so that $c(t_1, t_2) = c^*(v)$ for some function c^*. The three models we described above can be shown to belong to the Archimedian family of distributions and to correspond to three parametric forms for the function $c^*(v)$. In Clayton's model, e.g., the cross-ratio function $c^*(v)$ is independent of (t_1, t_2) and has the constant value θ. On the other hand, for stable model $c^*(v) = 1 + (1 - \omega)/(-\omega \log v)$ (Oakes, 1989), which is an increasing function of v, it approaches one as $v \to 0$ and approaches ∞ as $v \to 1$. Thus, in this model, for fixed v, the strength of association increases with ω decreasing, whereas for fixed ω, there is a stronger association for larger values of v (or smaller values of t_1 and t_2). In Frank's model, $c^*(v) = -v \log \kappa \{1 + \kappa^v/(1 - \kappa^v)\}$, which is also an increasing function of v but is linear in shape, with $c^* \to 1$ as $v \to 0$.

Even though the copula parameters for the parametric association models give a valid measure of association, they can be abstract quantities to interpret. Kendall's tau is a better understood global measure of association (Oakes, 1989) and, in this article, we often report the value of Kendall's tau corresponding to estimates of the copula parameters. Since the copula parameter has different interpretations in different models, Kendall's tau is also useful for comparing association estimates between the models. For the rest of the article, we will use α as a generic notation for the copula parameter irrespective of the choice of the model.

In our modeling and estimation approach, we also implicitly assume that, for a susceptible individual, the marginal distribution of his/her failure time does not depend on the susceptibility status of the other member, i.e.,

$$\begin{aligned} &\Pr(T_j^* \geq t_j \mid Y_j = 1, Y_i, i \neq j) \\ &\quad = \Pr(T_j^* \geq t_j \mid Y_j = 1), \qquad j = 1, 2. \end{aligned} \tag{2}$$

This assumption is similar to the so-called subject-specific-effect or reproducibility assumption often made in the marginal regression modeling approach for multivariate outcomes (cf., Whittemore, 1995). To see that Assumption 2 is needed, first note that, by marginalizing the joint distribution given in equation (1), we obtain,

$$S_j(t_j) = \Pr(T_j^* \geq t_j \mid Y_j = 1, Y_i = 1, i \neq j).$$

Since by definition of our marginal model $S_j(t) = \Pr(T_j^* \geq t \mid Y_j = 1)$, it follows that, for consistency of model definition, we need the assumption given in equation (2).

2.2 *Estimation*

Here we consider a setting suitable for the WAS example. The methods described in this section, however, are relevant for other applications of the proposed model as well.

For the ith of m families, let (δ_{ij}, t_{ij}), $j = 1, \ldots, n_i$, be the observed data on the history of a disease of n_i members of the family. Here δ denotes the indicator of whether the individual developed the disease ($\delta = 1$) or not ($\delta = 0$) during the follow-up and t denotes the integer age at onset (t^*) if $\delta = 1$ and the follow-up time (censoring time) if $\delta = 0$. Clearly, for an individual, $\delta = 1$ implies $y = 1$, i.e., the individual is susceptible. For $\delta = 0$, however, y is unknown due to censoring. Individuals within a family are assumed to have a common marginal distribution, described by the proportion of susceptible individuals, ϕ, and the distribution of the ages at onset among the susceptible individuals, $S(t)$. We assume that the joint distribution of any pair of individuals within a family can be described by the bivariate model we proposed, with the association parameters γ and α being common to all the pairs. No assumptions are made on third or higher order dependency structures. Finally, we make the following independent censoring assumptions: the joint susceptibility status of a pair of individuals is independent of their joint censoring times; given one member is susceptible and the other is not, the failure time of the susceptible individual is independent of the corresponding censoring time; and given both members are susceptible, their joint failure times are independent of the joint censoring times.

Now we construct a quasi-likelihood of the data under the proposed model that ignores the dependency among different pairs within the same family. The contribution of a pair of relatives is obtained from the likelihood of the paired data under the proposed model and the contribution of a family is obtained by taking the product over the contributions of all the possible pairs from that family. Thus, the quasi-likelihood has the form

$$L_2 = \prod_{i=1}^{m} L_{2i} = \prod_{i=1}^{m} \prod_{(j,k) \in C_i} L_{(j,k)_i}, \tag{3}$$

where C_i is the collection of all the possible pairs for family i and $L_{(j,k)_i}$ denotes the contribution of the pair consisting of the jth and kth member of the ith family. It follows that, under the proposed model and independent censoring assumptions, the likelihood of the data for a pair of relatives, up to a constant, is given by

$$\begin{aligned} L_{(j,k)_i} = {} & H_1(u_{ij}, u_{ik})^{\delta_{ij}\delta_{ik}} H_2(u_{ij}, u_{ik})^{(1-\delta_{ij})\delta_{ik}} \\ & \times H_3(u_{ij}, u_{ik})^{\delta_{ij}(1-\delta_{ik})} H_4(u_{ij}, u_{ik})^{(1-\delta_{ij})(1-\delta_{ik})}, \end{aligned} \tag{4}$$

where

$$\begin{aligned} H_1(u_{ij}, u_{ik}) &= \left\{ C_\alpha(u_{ij}^-, u_{ik}^-) - C_\alpha(u_{ij}^-, u_{ik}) - C_\alpha(u_{ij}, u_{ik}^-) \right. \\ &\qquad \left. + C_\alpha(u_{ij}, u_{ik}) \right\} p_{11}, \\ H_2(u_{ij}, u_{ik}) &= \left\{ C_\alpha(u_{ij}, u_{ik}^-) - C_\alpha(u_{ij}, u_{ik}) \right\} p_{11} \\ &\quad + (u_{ik}^- - u_{ik}) p_{01}, \\ H_3(u_{ij}, u_{ik}) &= \left\{ C_\alpha(u_{ij}^-, u_{ik}) - C_\alpha(u_{ij}, u_{ik}) \right\} p_{11} \\ &\quad + (u_{ij}^- - u_{ij}) p_{10}, \\ H_4(u_{ij}, u_{ik}) &= C_\alpha(u_{ij}, u_{ik}) p_{11} + u_{ij} p_{10} + u_{ik} p_{01} + p_{00}, \\ u_{ij} &= S(t_{ij}) \quad \text{and} \quad u_{ij}^- = S(t_{ij} - 1), \\ u_{ik} &= S(t_{ik}) \quad \text{and} \quad u_{ik}^- = S(t_{ik} - 1). \end{aligned}$$

Note that the contribution of each pair is given by a mixture of one, two, or four terms depending on whether none, exactly one, or both members of the pair are censored. If the first member's event was observed but the second member was censored, e.g., then their contribution to the quasi-likelihood is a mixture of two components, denoted by H_3 above, where the first component refers to the condition that both members are susceptible and the second component refers to the condition that only the first member is susceptible. We also note that we have assumed age to be discrete and to have been recorded to the nearest integer. These formulae can be easily modified for continuous age by replacing the first- and second-order differences by appropriate first-order and second-order derivatives of the joint survivor function.

The motivation for considering the above quasi-likelihood approach instead of the full likelihood of the data merits some discussion. Both measures of familial association we consider, namely γ and α, are marginal association parameters that characterize the association between two relatives at a time. The quasi-likelihood approach requires model specification for two relatives and treats higher order association as nuisance. The full likelihood approach, on the other hand, requires specification of a model for joint distribution of all the relatives in a family, which may not be desirable primarily for two reasons: complexity and sensitivity of the resulting inference. In principle, both the susceptibility model and the age-at-onset model can be extended to more than two relatives without changing the marginal interpretation of the parameters γ and α. For the age-at-onset model, the copula distribution can be easily extended to arbitrary family size if α can be assumed to be constant across pairs of individuals. But extension of the susceptibility model to large families will involve specifying the joint distribution for a large number of correlated binary random variables, which is typically a complex task (cf., Liang et al. (1992) for a discussion). Further complexity arises due to the latent structure of the problem. Since the susceptibility status is not observable for censored individuals, after the joint distributions for the age at onset and susceptibility are specified, they need to be mixed over all possible combinations of the susceptibility status of the censored family members. As the number of possible combinations increases geometrically with the family size, the full likelihood becomes increasingly complex. In our study, e.g., a significant number of families have five or more members. For these families, computation of the likelihood may involve mixing over $2^5 = 32$ or more terms. Finally, we note that the consistency of the marginal parameters of interest in the full likelihood approach relies on the correct specification of the full joint distribution, which is more stringent than the assumption of correct specification of the model for pairs of individuals, a condition sufficient for the consistency of the parameter estimates from the quasi-likelihood.

2.3 *Two-Stage Parametric Estimation*

Assume $S(t)$ has a known parametric form $S(t;\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ denotes a vector of parameters whose values are unknown. At stage 1 of the two-stage estimation method, we consider estimating the marginal parameters $\boldsymbol{\beta}$ and ϕ using a univariate mixture analysis approach that ignores the dependency of the individuals within a family. The estimates of $\boldsymbol{\beta}$ and ϕ are obtained by maximizing the following marginal likelihood:

$$\begin{aligned} L_1 &= \prod_{i=1}^{m} L_{1i} \\ &= \prod_{i=1}^{m}\prod_{j=1}^{n_i} \left[\phi\left\{S(t_{ij}-1;\boldsymbol{\beta}) - S(t_{ij};\boldsymbol{\beta})\right\}\right]^{\delta_{ij}} \\ &\qquad \times \left\{\phi S(t_{ij};\boldsymbol{\beta}) + 1 - \phi\right\}^{1-\delta_{ij}}. \end{aligned} \tag{5}$$

In (5), the contribution from a censored individual is obtained as a mixture because of the unknown susceptibility status of the individual. It is well known that such a marginal likelihood approach that ignores the correlation between individuals gives consistent estimates of the parameters of the marginal distributions and that the estimates are robust against the misspecification of the correlation structure. At stage 2, estimates of the association parameters are obtained by fixing the marginal distributions at their estimates and maximizing the quasi-likelihood (3) with respect to only α and γ.

Some further notation is useful in deriving asymptotic theory of the estimates. The joint estimating equation for $\psi = (\phi, \boldsymbol{\beta})'$ and $\eta = (\gamma, \alpha)'$ can be written as two sets of equations,

$$\sum_{i=1}^{m} U_{1i}(\psi) = 0 \quad \text{and} \quad \sum_{i=1}^{m} U_{2i}(\psi, \eta) = 0, \tag{6}$$

where $U_{1i}(\psi) = \Sigma_{j=1}^{n_i} \partial/\partial\psi \log L_{1i}$ and $U_{2i}(\psi,\eta) = \Sigma_{j=1}^{n_i} \partial/\partial\eta \log L_{2i}$. The asymptotic property of $(\hat{\psi}, \hat{\eta})$ can be stated as follows.

THEOREM 1: *Assuming that the true joint distribution for pairs of relatives satisfies the specified parametric model for some true parameter values* (ψ_0, η_0) *belonging to the interior of the parameter space, as* m *(number of clusters)* $\rightarrow \infty$, $(\hat{\psi}', \hat{\eta}')$ *are consistent for* (ψ_0', η_0') *and* $m^{1/2}[(\hat{\psi} - \psi_0)', (\hat{\eta} - \eta_0)']$ *converges to a multivariate normal distribution with mean zero and covariance matrix* $A^{-1}BA^{-1}$, *where*

$$A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix},$$

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{12}^T & B_{22} \end{bmatrix},$$

with

$$A_{11} = -\lim_{m\to\infty} m^{-1} \frac{\partial}{\partial\psi'} \sum_{i=1}^{m} U_{1i}(\psi)|_{(\psi=\psi_0)},$$

$$A_{21} = -\lim_{m\to\infty} m^{-1} \frac{\partial}{\partial\psi'} \sum_{i=1}^{m} U_{2i}(\psi,\eta)|_{(\psi=\psi_0,\eta=\eta_0)},$$

$$A_{22} = -\lim_{m\to\infty} m^{-1} \frac{\partial}{\partial\eta'} \sum_{i=1}^{m} U_{2i}(\psi,\eta)|_{(\psi=\psi_0,\eta=\eta_0)},$$

and

$$B_{11} = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \operatorname{var}\{U_{1i}(\psi_0)\},$$

$$B_{12} = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \operatorname{cov}\{U_{1i}(\psi_0), U_{2i}(\psi_0, \eta_0)\},$$

$$B_{22} = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \operatorname{var}\{U_{2i}(\psi_0, \eta_0)\}.$$

The asymptotic covariance matrix can be consistently estimated by replacing (ψ_0, η_0) by their estimates and the means and variances by their corresponding empirical counterparts. The proof of the asymptotic normality theorem follows by using a standard first-order Taylor's series expansion of the joint estimating equation given in (6). It follows easily that $m^{1/2}[(\hat{\psi} - \psi_0)', (\hat{\eta} - \eta_0)']$ is asymptotically equivalent to $1/m^{1/2} A^{-1} \Sigma_{i=1}^{m} [U_{1i}(\psi_0), U_{2i}(\psi_0, \eta_0)]'$, which is the sum of m independent (may not be identically distributed) terms. Finally, an application of Liapounov's central limit theorem for the sum of independently but not necessarily identically distributed random variables yields the desired result.

2.4 *Two-Stage Semiparametric Estimation*

Here we consider estimating the marginal distribution nonparametrically at the first stage and estimating the association parameters at the second stage with the marginal distribution fixed at the corresponding nonparametric estimate. We propose estimating the marginal lifetime risk (ϕ) using the maximum value of the marginal Kaplan–Meier empirical distribution function based on (t_{ij}, δ_{ij}), $j = 1, \ldots, n_i$, $i = 1, \ldots, m$. Thus, if T_N, where $N = \Sigma n_i$, denotes the maximum of the t_{ij}'s and $\hat{K}(t)$ denotes the Kaplan–Meier distribution function, we have $\hat{\phi} = \hat{K}(T_N)$. The cumulative distribution $F(t) = 1 - S(t)$ can be estimated by normalizing the improper distribution function $\hat{K}(t)$ as $\hat{F}(t) = \hat{K(t)}/\hat{\phi}$. For independent and identically distributed censored failure time data, Maller and Zhou (1992, 1995) developed asymptotic properties of such nonparametric estimators. Further work is needed to derive the asymptotic properties of these estimators in a correlated data context and hence obtain the asymptotic variances for the semiparametric estimates of the association parameters in our model. Alternatively, one can consider a bootstrap approach with families as the bootstrap sampling units to obtain the variance of the semiparametric estimator. The bootstrap, however, needs to be properly done so that it reflects the true uncertainty of the estimates. In the following section, we describe the bootstrap procedure for our specific example.

3. Example

We applied the proposed methodology to the WAS data. Information on age at onset of breast cancer for the female first-degree relatives of the volunteers is available to us. The relatives are treated as a cohort of individuals who are followed from their birth until the incidence of cancer or the censoring time. The ages of the relatives at the time of the interview of the volunteer or, for deceased relatives, the age at which the relative died defines the censoring times. A total of 13,223 subjects coming from 4856 distinct families was used to estimate the marginal survival. For the parametric marginal distribution, we used the Weibull model with the form $S(t; \boldsymbol{\beta}) = \exp\{-(\beta_1 t)^{\beta_2}\}$.

Table 1
Frequencies of number of WAS families by the number of pairs a family contribute

Number of pairs/family	Number of families
1	1858
2	385
3	526
4	217
≥5	224

In estimating the association parameters, we chose a subset of the data in which the subjects in the same family are first-degree relatives to each other. It was felt that the pairwise association should be similar if the members of the pairs are first-degree relatives to each other. This subset of data contains 6769 pairs coming from 3210 families. Table 1 is a tally of the number of such pairs each family contributes to the estimation of the association parameters. More than half of the families contribute only one pair, and the majority of the families contribute no more than three pairs.

To obtain the bootstrap variance of the semiparametric estimates, we draw simple random samples with replacement from the 4856 families of the same size as total number of families. For each of the bootstrap samples of families, we apply our two-stage estimation procedure. The first stage involves all the individuals in the selected families, but the second stage involves only a subset of these individuals; only those families that have at least two family members are included and then, from each such family, we construct all possible pairs of relatives so that the members of the pairs are first-degree relatives to each other. The bootstrap variance estimate is obtained as the variance of the parameter estimates from 2000 bootstrap samples.

Figure 1 displays the nonparametric and parametric estimate of $\phi \times F(t)$, the cumulative risk of breast cancer for a woman with unknown susceptibility status. Close agreement

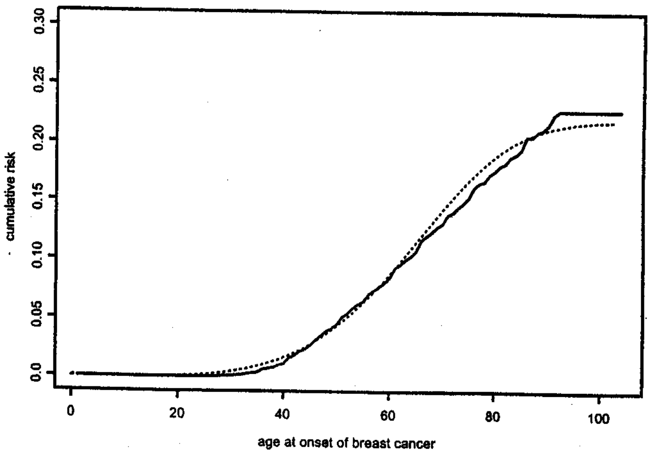

Figure 1. Nonparametric and parametric estimates of the cumulative risk of breast cancer from the WAS study. The solid line is the nonparametric estimate and the dotted line is the parametric estimate.

Table 2
Estimates (SE) from the WAS study

		Parametric margins $S_0(t) = \exp\{-(\beta_1 t)^{\beta_2}\}$	
$\log \beta_1$		−4.24 (.0156)	
$\log \beta_2$		1.53 (.0216)	
ϕ		0.22 (.0093)	
	Clayton	Frank	Stable
γ	2.76 (.5023)	2.79 (.5384)	2.84 (.5486)
α	$\theta = 1.39$ (.2085)	$\kappa = 0.24$ (.0634)	$\omega = 0.84$ (.0547)
		Nonparametric margins	
ϕ		0.23 (.0103)	
	Clayton	Frank	Stable
γ	2.67 (.6209)	2.84 (.4652)	2.94 (.5025)
α	$\theta = 1.51$ (.2821)	$\kappa = 0.19$ (.1212)	$\omega = 0.89$ (.0450)

of the nonparametric and parametric curves indicates that the Weibull model fits the data well. The oldest incidence occurred at age 91, after which the Kaplan–Meier cumulative risk estimate leveled off at 0.23. Between age 91 and the last observation at age 103, there were 178 censored observations. Such a large number of subjects censored at the very old ages provides strong evidence that the Kaplan–Meier curve has truly leveled, and the leveled value from the Kaplan–Meier should be a good estimate of the lifetime risk.

Parametric estimates and their standard errors (the standard errors for the semiparametric estimation were bootstrap standard errors) are summarized in Table 2. For modeling the association between age at onset of relatives, we considered three models described in Section 2.1. We note that, due to the use of the marginal likelihood approach at stage 1, estimates of the parameters of the marginal distribution do not depend on the choice of these association models. Both the estimate and standard error of ϕ are remarkably similar in the parametric and nonparametric models. Irrespective of the choice of nonparametric or parametric marginal distribution and the choice of the copula models, estimates of γ show a strong and significant association between susceptibilities of relatives. The copula parameter has different interpretations in different models and hence is not directly comparable between the models. However, the parameter estimates from all the models considered correspond only to weak association. For the parametric estimate of $\kappa = 0.24$, e.g., Kendall's tau is only 0.18. In general, Kendall's tau corresponding to the various estimates ranged between 0.1 and 0.2. Thus, we see that familial association is mostly seen in joint susceptibility, which suggests that the combined effect of various familial risk factors for breast cancer in this population is expressed mostly in terms of overall susceptibility. Finally, we note that, if the possibility of correlation in overall susceptibility was left out, i.e., if γ was fixed at 1.0, then the semiparametric estimate for each of θ, κ, and ω would correspond to a much stronger association for joint age at onset (data not shown), which results from absorbing the association from the joint susceptibility into the association between ages at onset.

4. Simulations

We conducted a simulation study to evaluate the two-stage estimator. We generated 5000 (mother, sister) pairs from the bivariate mixture model with parameters similar to those estimated from the semiparametric model. We chose Clayton's association model for the joint age-at-onset distribution. Censoring times for the mothers were generated from N(75, 15) and for the sisters were generated from N(50, 12), both of which were similar to those observed in the WAS data. The simulation experiment was repeated 250 times.

Table 3 presents the simulation results. Both parametric and semiparametric approaches produced little bias. The large variances for the association parameters γ and θ are mainly due to the small incidence rate. With $\phi = .2$ and censoring, the proportion of incidence observed is less than 10% on average. Thus, the joint incidence is sparse. When ϕ is increased to 0.3, there is a substantial reduction of variances for γ and θ. It is interesting to note that the semiparametric approach has about the same efficiency as the parametric procedure. In particular, relaxing the parametric assumption on the marginal distribution does not appear to reduce the efficiency of the association parameters. We see that, when $\gamma = 1$ is forced, i.e., only correlation between ages at onset is allowed and correlation between susceptibilities are ignored (which is equivalent to the standard practice of modeling the joint age at onset by the copula models themselves), the semiparametric estimate for θ gives an overestimate of the true correlation between the ages at onset. Finally, Table 3 also shows the performance of the sandwich variance estimator we proposed in Section 2.2 for two-stage parametric estimation. By comparing the means of the estimated standard errors over the simulated data sets (fourth column) to the empirical standard errors (third column), we see that, overall, the proposed sandwich standard estimator performs well in estimating the true standard errors.

5. Discussion

In this article, we have proposed a bivariate distribution for flexible modeling of correlation in familial disease data. Two

Table 3

Performance of the two-stage parametric and semiparametric estimators in Clayton's copula model from 250 *samples of* 5000 *pairs*

	Mean	SE	Mean($\widehat{\text{SE}}$)		Mean	SE	Mean($\widehat{\text{SE}}$)
Parametric Margins							
$\phi = .2$	.20	.007	.007	$\phi = .3$	.29	.008	.008
$\log \beta_1 = -4.23$	-4.23	.012	.011	$\log \beta_1 = -4.23$	-4.23	.009	.009
$\log \beta_2 = 1.58$	1.58	.029	.028	$\log \beta_2 = 1.58$	1.58	.023	.023
$\gamma = 2.72$	2.81	.628	.629	$\gamma = 2.72$	2.79	.534	.549
$\theta = 1.50$	1.55	.324	.319	$\theta = 1.50$	1.51	.201	.212
Nonparametric Margins							
	Mean	SE			Mean	SE	
$\phi = .2$	.20	.007		$\phi = .3$	.29	.009	
$\gamma = 2.72$	2.82	.625		$\gamma = 2.72$	2.78	.534	
$\theta = 1.50$	1.54	.318		$\theta = 1.50$	1.50	.203	
Nonparametric Margins, Correlation in Ages at Onset Only							
	Mean	SE			Mean	SE	
$\theta = 1.50$	2.21	.282		$\theta = 1.50$	1.98	.183	

types of expression for the disease are considered simultaneously, the disease incidence and the age at onset of the diseased individuals. Odds ratio modeling is considered for measuring association between two individuals in terms of their disease incidence, whereas parametric correlation models induced by the copula class of functions are considered for modeling correlation between ages at onset of diseased individuals. The model can be easily extended to incorporate covariate information, both into the marginal and the association parameters. Typically, for studying familial diseases, covariates for the marginal parameters include the observable risk factors of the disease and the covariates for the association parameters include the relationships between relatives. It may be of scientific interest to see whether the incidence and the age at onset of the disease depend on different sets of covariates.

In a univariate setting, mixture models have received considerable attention in various areas of survival analysis, including analysis of data on time to recurrence (or death) of diseases (Boag, 1949; Gordon, 1990), analysis of long-term survivorship in toxicological animal experiments (Farewell, 1982; Taylor, 1995), and study of the incubation period of AIDS after HIV infection (Lui, Darrow, and Rutherford, 1988; Struthers and Farewell, 1989). A common theme in all these areas of application is that a fraction of individuals has been assumed to be cured or immuned so that they can never observe the end-point event. Thus, in this approach, individuals who are considered censored in an ordinary survival analysis are considered to be a mixture of the cured individuals and the noncured individuals who are censored due to incomplete follow-up. Maller and Zhou (1996) is an excellent reference for this literature. The bivariate mixture model we propose can be thought of as an extension of the univariate cure models to the bivariate situation. Potentially, our model will be useful in analysis of correlated survival data when cure is a possibility. Our model, e.g., can be used to study familial association in disease recurrence after treatment. For such data, it is likely that some proportion of individuals are cured by the treatment and will never experience the disease. Estimating the correlation between cure probabilities between related individuals can shed light on possible interactions between the treatment and genetic factors. Similarly, in toxicological animal studies, the proposed model can be used to investigate possible litter effect on immunity of animals to the toxicant. Models for bivariate survival data that assume correlations among individuals are expressed only in terms of time to event may not be satisfactory for such data.

As in the case of univariate mixture models, there may be situations when the bivariate model faces lack of identifiability or/and interpretability. Consider a hypothetical situation where everybody will eventually have a particular disease if they are alive for a very long time but we only observe a small fraction of the disease incidence because individuals have a feasible maximal lifespan due to other causes of death, say about 100 years. It seems that the lifetime risk in this situation just gives an estimate of the cumulative probability of the disease (marginal for univariate models and joint for bivariate models) until 100 years. Thus, we see that interpretation of the overall susceptibility or lifetime risk of individuals depends on the extremum of the distribution of the survival time of patients from other causes of death (censoring time). Since very few people can be expected to live up to the maximal lifespan, it also seems that there is very little information in estimating cumulative risk until 100 years. We note that, in the real example considered in this article, the oldest observed incidence of the disease was 91, even though a relatively large number of individuals were followed beyond that age. This suggests that, in this population, the probability of developing breast cancer after age 91 is very small and hence the estimate of cumulative risk we obtain till age 91 is a good approximation of the lifetime risk. For proper interpretation of parameters, it can be important to establish whether the hypothesized populations of insusceptible, immuned, or cured individuals truly exist or not. For i.i.d data, Maller and Zhou

(1992, 1996) have developed formal methods of testing for the existence of such a population based on the properties of the Kaplan–Meyer estimator of the cumulative risk function. Future theoretical work is needed to extend this method in the context of correlated data. A starting point of this research could be the work by Ying and Wei (1994) on the properties of the Kaplan–Meyer estimator for dependent data.

RÉSUMÉ

Pour modéliser la corrélation dans les maladies familiales avec la variable "age de début", nous proposons un modèle bivarié, qui incorpore deux types d'associations par paires, l'une entre le risque lié à la durée de vie ou susceptibilité globale de deux individus, et l'autre entre les ages de début chez deux individus susceptibles. Pour l'estimation, nous considérons une procédure à deux étapes similaire à celle de Schih (1998). Nous évaluons les propriétés des estimateurs au moyen de simulations et comparons la performance avec celle d'un modèle de survie bivarié qui permet une corrélation entre les ages de début seulement. Nous appliquons la méthodologie au cancer du sein en utilisant les données de parenté de l'étude "Washington Ashkenazi". Nous discutons également les applications potentielles de la méthode proposée dans le domaine de la modélisation de la guérison.

REFERENCES

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B* **11,** 15–44.

Ciske, D. J., Stephen, S. R., King, R. A., Anderson, V. E., Bartow, S., Vachon, C., McGovern, P. G., Kushi, L. H., Zheng, W., and Sellers, T. A. (1996). Segregation analysis of breast cancer: A comparison of type-dependent age-at-onset versus type-dependent susceptibility models. *Genetic Epidemiology* **13,** 317–328.

Claus, E. B., Risch, N., and Thompson, W. D. (1991). Genetic analysis of breast cancer in the Cancer and Steroid Hormone Study. *American Journal of Human Genetics* **48,** 232–242.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65,** 141–151.

Elston, R. C. and Yelverton (1975). General models for segregation analysis. *Annals of Human Genetics* **27,** 31–45.

Farewell, V. T. (1977). A model for binary variable with time censored observations. *Biometrika* **64,** 43–46.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long term survivors. *Biometrics* **38,** 1041–1046.

Farewell, V. T., Math, B., and Math, M. (1977). The combined effect of breast cancer risk factors. *Cancer* **40,** 931–936.

Frank, M. J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematicae* **19,** 194–226.

Genest, C. and MacKay, R. J. (1986). Archimedian copulas and bivariate families with continuous marginals. *American Statistician* **40,** 280–283.

Gordon, N. H. (1990). Application of the theory of finite mixtures for the estimation of 'cure' rates of treated cancer patients. *Statistics in Medicine* **9,** 397–407.

Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika* **73,** 671–678.

Kuk, A. Y. C. and Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazard regression. *Biometrika* **79,** 531–541.

Liang, K. Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B* **54,** 3–40.

Lui, K.-J., Darrow, W. W., and Rutherford, G. W. (1988). A model-based estimate of the mean incubation period for AIDS in homosexual men. *Science* **240,** 1333–1335.

Maller, R. A. and Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika* **79,** 731–739.

Maller, R. A. and Zhou, S. (1995). Testing for the presence of immune or cured individuals in censored survival data. *Biometrics* **51,** 1197–1205.

Maller, R. A. and Zhou, S. (1996). *Survival Analysis with Long-Term Survivors.* Chichester, U.K.: John Wiley and Sons.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84,** 487–493.

Shih, J. H. (1998). Modeling multivariate discrete failure time data. *Biometrics* **54,** 1115–1128.

Struewing, J. P., Hartge, P., Wacholder, S., Baker, S. M., Berlin, M., McAdams, M., Timmerman, M. M., Lawrence, B. C., and Tucker, M. A. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *The New England Journal of Medicine* **336,** 1401–1408.

Struthers, C. A. and Farewell, V. T. (1989). A mixture model for time to AIDS data with left truncation and an uncertain origin. *Biometrika* **76,** 814–817.

Taylor, J. M. G. (1995). Semiparametric estimation in failure time mixture models. *Biometrics* **51,** 899–907.

Whittemore, A. S. (1995). Logistic regression of family data from case–control studies. *Biometrika* **82,** 57–67.

Ying, Z. and Wei, L. J. (1994). The Kaplan–Meier estimate for dependent failure time observations. *Journal of Multivariate Analysis* **50,** 17–29.

Zhao, L. and Prentice, R. (1992). Correlated binary data using quadratic exponential model. *Biometrika* **77,** 642–648.

Received May 2000. *Revised March* 2001.
Accepted March 2001.